AudioGPT: Understanding and Generating Speech, Music, Sound, and Talking Head

Year: 2023 | Citations: 317 | Authors: Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang

---

Abstract

Large language models (LLMs) have exhibited remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. Despite the recent success, current LLMs are not capable of processing complex audio information or conducting spoken conversations (like Siri or Alexa). In this work, we propose a multi-modal AI system named AudioGPT, which complements LLMs (i.e., ChatGPT) with 1) foundation models to process complex audio information and solve numerous understanding and generation tasks; and 2) the input/output interface (ASR, TTS) to support spoken dialogue. With an increasing demand to evaluate multi-modal LLMs of human intention understanding and cooperation with foundation models, we outline the principles and processes and test AudioGPT in terms of consistency, capability, and robustness. Experimental results demonstrate the capabilities of AudioGPT in solving 16 AI tasks with speech, music, sound, and talking head understanding and generation in multi-round dialogues, which empower humans to create rich and diverse audio content with unprecedented ease. Code can be found in https://github.com/AIGC-Audio/AudioGPT